XML Based System for Corpus Linguistics

CL@RK System

SFS Tuebingen - LML Sofia

Etude de CL@RK System

Par
Henry Fabiola
&
Lordonnois Claire

Introduction

- CL@RK System est un logiciel né d ’un projet universitaire conjoint de la SFS de Tuebingen en Allemagne et de la LML de Sofia en Bulgarie.

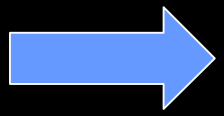
- Au départ, le projet a été créé pour définir un modèle d'apprentissage du Bulgare via une grammaire. Le logiciel a pour but principal de faciliter la constitution de corpus.

- Clark Development team :

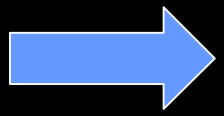
info@bultreebank.org

CL@RK System un outil TAL

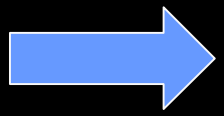
- est un éditeur XML qui permet avant tout la gestion et la manipulation de corpus.



Constitution de corpus



étiquetage de corpus



édition de corpus

CL@RK System version 1.0

- La version complète et gratuite est téléchargeable en 1 mn avec une connection ADSL sur le site : http://www.BulTreeBank.org/clark/index.html

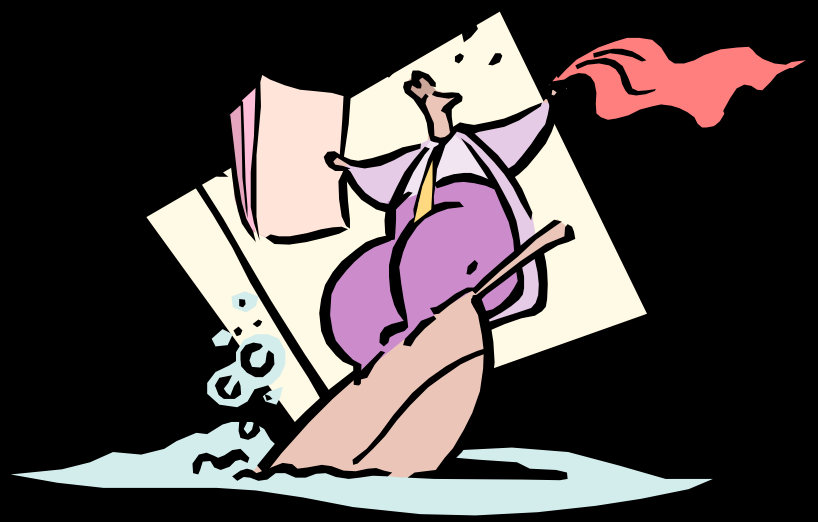
- Elle est également disponible sur le CD TAL1 dans OUTIL TAL

Exigences d'installation

- CL@RK System recquiert la présence de JAVA 1.3.1 au minimum.
- JAVA est téléchargeable gratuitement sur le site : http://java.sun.com/
- CL@RK System est régi par une licence Free Software :

  Copyright: Kiril Simanov, Alexander Simov, Milen Kouylekov, Krasimira Ivanova

L'aide on line et off line

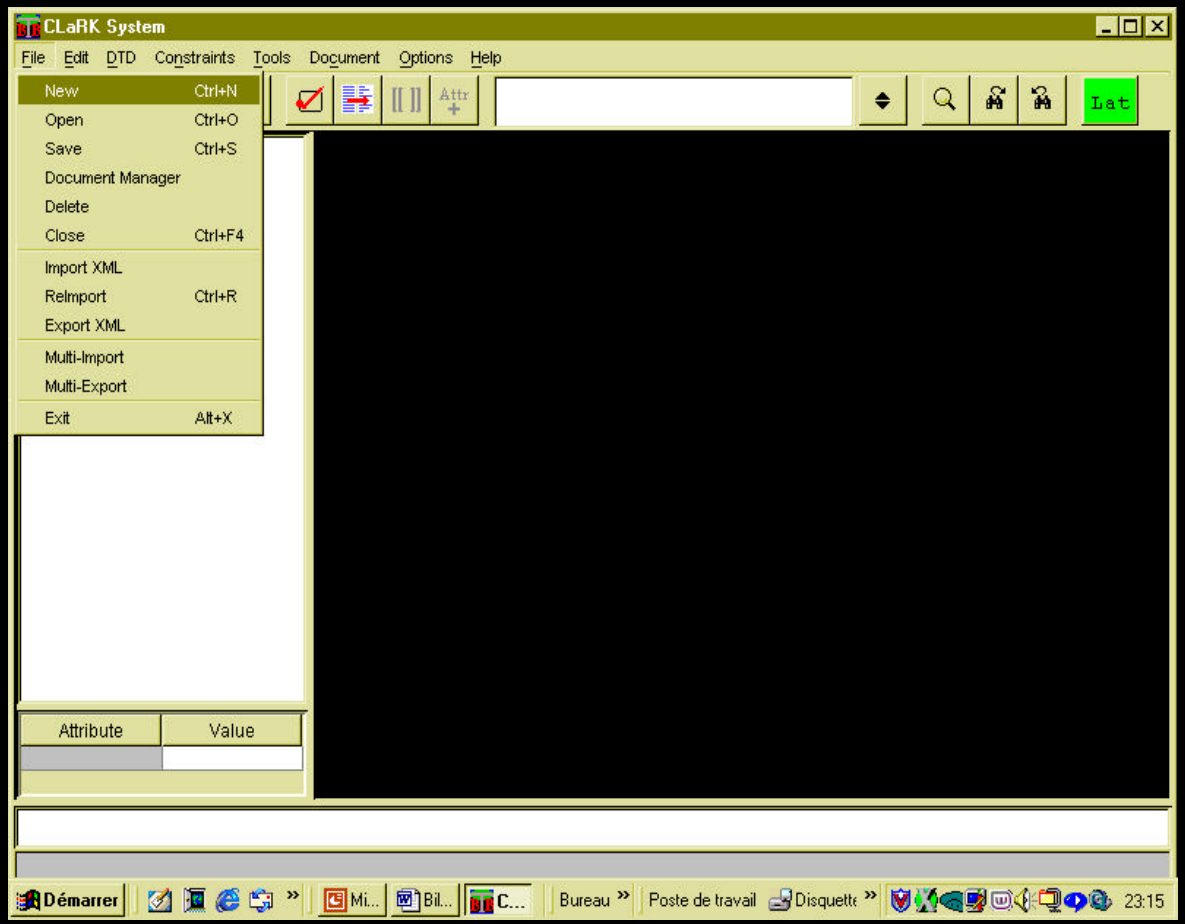


- Le manuel d'utilisation de CL@RK System au format PDF ou HTML est compris dans le package lors du téléchargement.
- Le manuel est également en ligne sur le site : http://www.BulTreeBank.org/clark/index.html
- Pour toute remarque, on peut s'adresser à : info@bultreebank.org

Les principaux sites parlant de CL@RK System

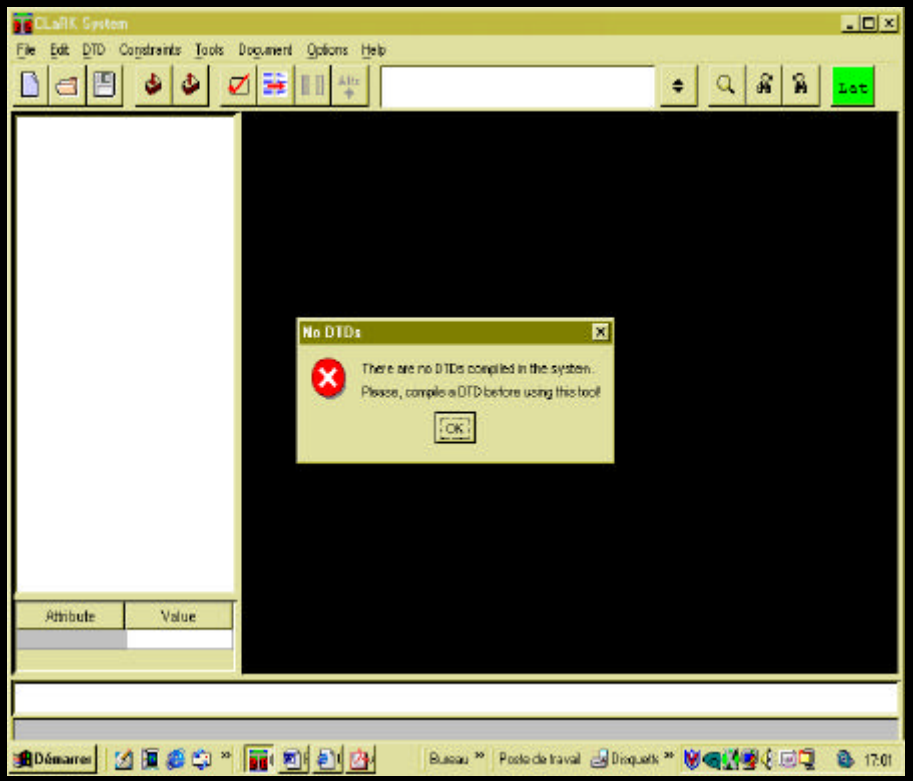
- CLaRK : http://www.sfs.nphil.uni-tuebingen.de/clark
- BulTreeBank : http://www.BulTreeBank.org

Un premier contact avec CL@RK System



Ouverture d ŭn nouveau document XML

- On ne peut pas ouvrir un nouveau document sans avoir préalablement déclaré la DTD à laquelle il est associé.
- CL@RK System édite a priori des documents valides.


File Edit DTD Constraints Tools Document Options Help
No DTDs
There are no DTDs compiled in the system.
Please, compile a DTD before using this tool!
OK
Attribute
Value

Edition de documents XML valides (1)

Le logiciel suit les recommandations de la W3C

VERIFICATION

Affichage d'un document valide sous un autre éditeur et un navigateur.

Edition de documents XML valides (2)

Pour vérifier que CL@RK System édite des document valides, nous avons :

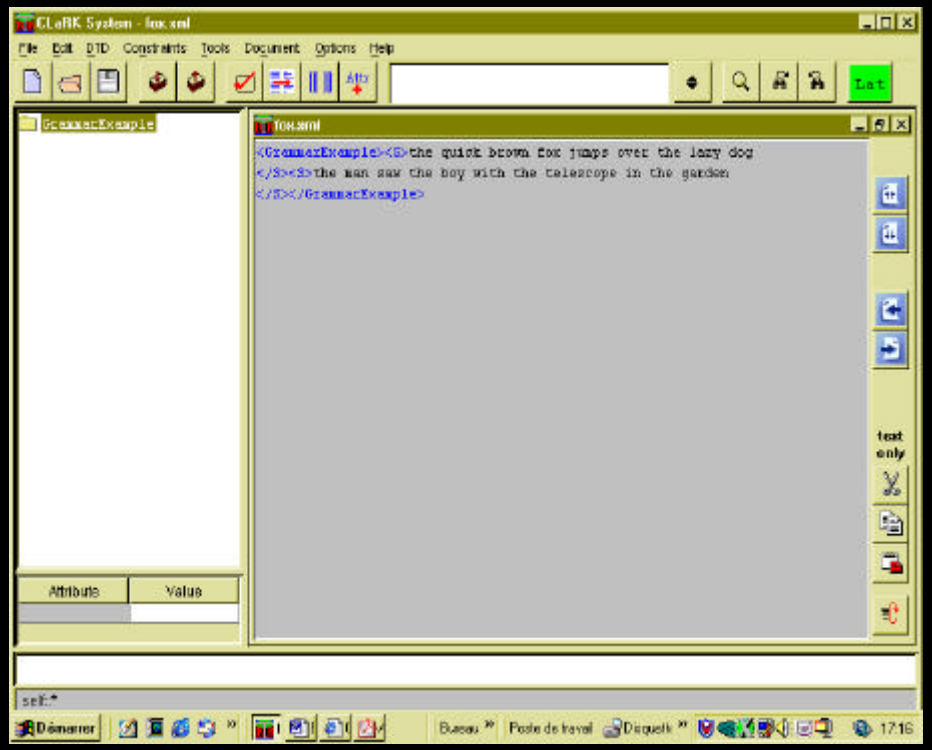
- Ouvert une DTD existante, grammar.dtd et son fichier associé, fox.xml, via les commandes :

  DTD : Menu/Compile DTD/grammar.dtd

  Fichier XML : Menu/File/Open/fox.xml

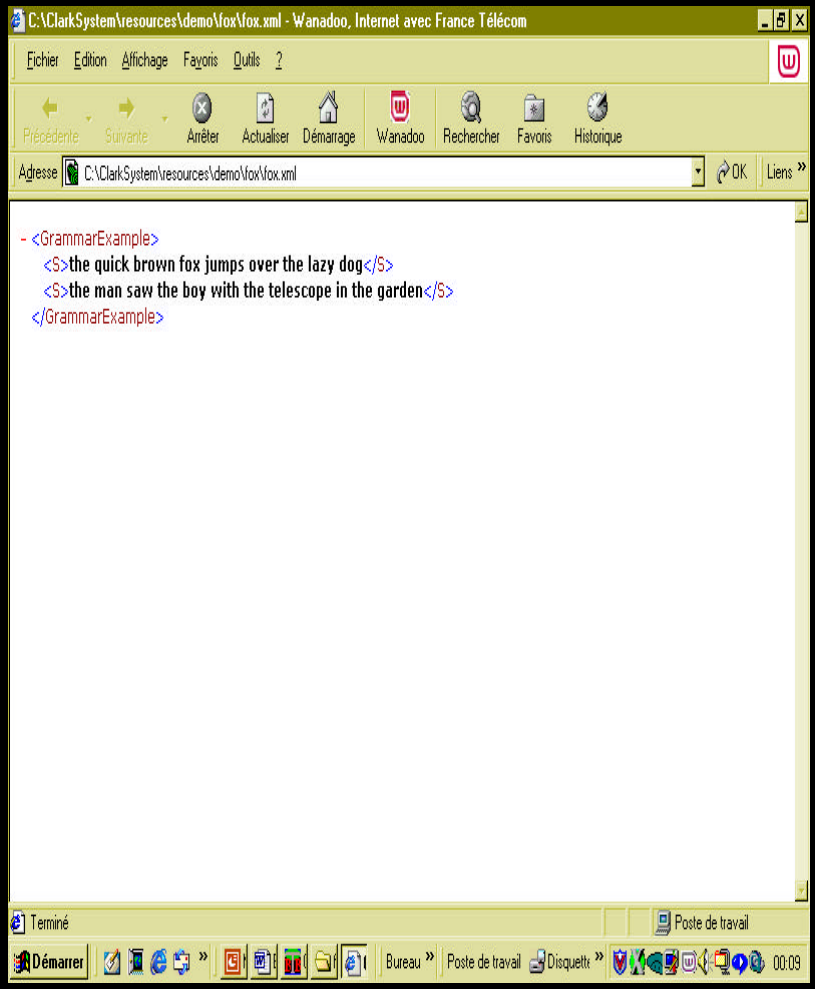
Edition de documents XML valides (3)

- A l'ouverture du fichier fox.xml, on constate l'abscence de prologue, <?xml version="1.0" encoding="ISO-8859-1"?> , et de balise indiquant que le document est associé à une DTD. Exemple : <!DOCTYPE texte SYSTEM "saintex.dtd">



Edition de documents XML valides (4)

- soumis Le fichier, fox.xml, à Internet Explorer pour vérifier sa validité.
- Conclusion : CL@RK System édite des fichiers XML valides.
- De plus, une partie du code est cachée à l utilisateur afin qu il puisse se concentrer sur le balisage du corpus.



Edition de DTD

- On ne peut pas créer directement de DTD sous CL@RK System comme on pourrait le faire avec XMLSPY.
- Une DTD peut être déjà existante, importée ou exportée.
- Les entités, les attributs, les sections CDATA et NMTOKEN sont gérés par une boîte de dialogue qui permet une visualisation des composants d'une DTD.
- Une conversion d'entités est automatique via Menu/Tools/Entity Converters.
- Seules les DTD externes sont traitées.

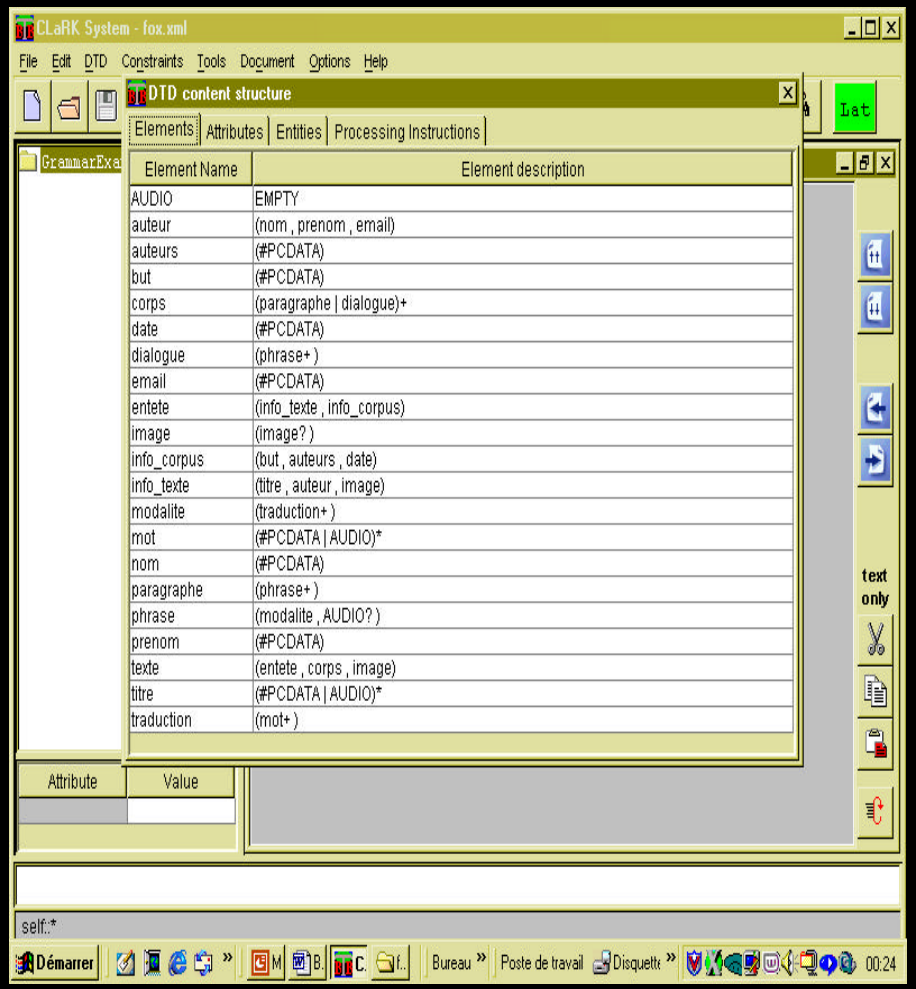
Visualisation d une DTD

- Importer/choisir une DTD :

  Menu/DTD/Compile DTD/fichier.dtd

- Visualiser la DTD :

  Menu/DTD/View DTD/fichier.dtd

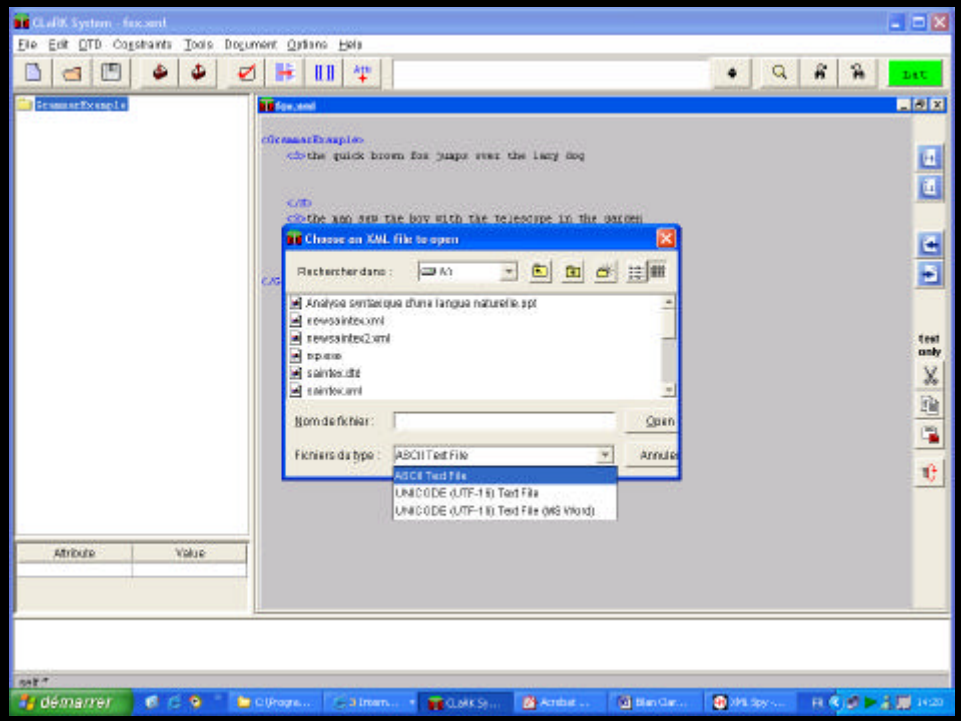


CL@RK System et UNICODE

Types d'encodage supportés:

ASCII (Europe et Europe de l'Ouest)

Unicode UTF 16 (langage universel)



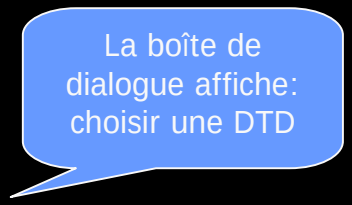
CL@RK System et XSL (1)

- CLaRK System possède une fonction de transformation XSLT qui fait également appel à Xpath.
- Comme tout document ouvert sous CLaRK, une DTD doit être assignée à un fichier.xsl avant qu'il soit créé.

CL@RK System et XSL (2)

Mode d'emploi:

File/Import XML/choose a file.



- La transformation peut être effectuée si le document validé est associé à une feuille de style uniquement interne. Le résultat est un nouveau document enregistré dans le système.

Quelques plus chez CL@RK System

- Utilisation d ėxpressions régulières : Menu/Contrainsts/
- Création de grammaires : Menu/Tools/Grammars
- Création de concordances : Menu/Tools/Concordance
- Commande de tri : Menu/Tools/Sort Tools
- Création de statistiques : Menu/Tools/Statistics

Une note pour CL@RK System

- Fonctionnalités XML
- Fonctionnalités secondaires
- Cohérence de l'interface
- Qualité graphique
- Clarté des messages
- Adaptabilité, flexibilité dans la personnalisation
- Demande confirmation avant opérations irréversibles
- Utilisabilité générale de l'interface
- Clarté et précision de la documentation

Conclusion

- CL@RK System est un bon outil pour traiter des corpus via XML.
- De nombreuses fonctions sont à disposition pour travailler sur les corpus, telles que : concordance, expressions régulières, et statistiques.
- CL@RK System est avant tout un analyseur syntaxique plus qu ́un éditeur XML comme Xml SPY.
- Les fonctions XSLT ne sont pas toujours faciles à manipuler.